Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

12-2023

Large language model is not a good few-shot information extractor, but a good reranker for hard samples!

Yubo MA

Yixin CAO
Singapore Management University, yxcao@smu.edu.sg

YongChin HONG

Aixin SUN

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

Part of the Databases and Information Systems Commons, and the Programming Languages and Compilers Commons

Citation

MA, Yubo; CAO, Yixin; HONG, YongChin; and SUN, Aixin. Large language model is not a good few-shot information extractor, but a good reranker for hard samples!. (2023). *The 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, December 6-10*.
Available at: https://ink.library.smu.edu.sg/sis_research/8388

This Conference Paper is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

arXiv:2303.08559v1 [cs.CL] 15 Mar 2023

Large Language Model Is Not a Good Few-shot Information *Extractor*, but a Good *Reranker* for Hard Samples!

Yubo Ma[1], **Yixin Cao**[2], **YongChing Hong**[1], **Aixin Sun**[1]

[1] S-Lab, Nanyang Technological University
[2] Singapore Management University
yubo001@e.ntu.edu.sg

Abstract

Large Language Models (LLMs) have made remarkable strides in various tasks. However, whether they are competitive few-shot solvers for information extraction (IE) tasks and surpass fine-tuned small Pre-trained Language Models (SLMs) remains an open problem. This paper aims to provide a thorough answer to this problem, and moreover, to explore an approach towards effective and economical IE systems that combine the strengths of LLMs and SLMs. Through extensive experiments on eight datasets across three IE tasks, we show that LLMs are not effective few-shot information extractors in general, given their unsatisfactory performance in most settings and the high latency and budget requirements. However, we demonstrate that LLMs can well complement SLMs and effectively solve *hard* samples that SLMs struggle with. Building on these findings, we propose an adaptive *filter-then-rerank* paradigm, in which SLMs act as filters and LLMs act as rerankers. By utilizing LLMs to rerank a small portion of difficult samples identified by SLMs, our preliminary system consistently achieves promising improvements (2.1% F1-gain on average) on various IE tasks, with acceptable cost of time and money.

1 Introduction

Large Language Models (LLMs, Brown et al. 2020; Chowdhery et al. 2022; Hoffmann et al. 2022) are becoming a fundamental tool for general task solver. They have shown amazing emergent (*e.g.,* memorization and reasoning) abilities through in-context learning (ICL) on various applications, including arithmetic reasoning, commonsense reasoning, and open-domain question answering (Wei et al., 2022c; Yu et al., 2023; Sun et al., 2023).

Recent studies have compared the performance between LLMs with ICL[1] and Small Language Models (SLMs) with conventional fine-tuning techniques[2] across many tasks. Focusing on information extraction (IE), some claim that LLMs are good few-shot extractors (Wang et al., 2022; Agrawal et al., 2022), while some others hold opposite opinions (Jimenez Gutierrez et al., 2022). We attribute the disagreement to the different IE subtasks, datasets, and settings in experiments. Given the disagreement, we claim a systematic evaluation on *whether LLMs perform competitively on various few-shot IE tasks* is non-trivial for further research.

In this paper, we target such a thorough evaluation on the advantages and disadvantages of LLMs and SLMs on various IE tasks. We aim to answer the following questions: **1)** Do LLMs really outperform SLMs in few-shot IE tasks? **2)** Can more annotations improve LLMs and SLMs? **3)** In terms of financial and time cost, which is preferable? **4)** Are LLMs and SLMs respectively adept at handling different types of samples?

To answer the first three questions, we conduct an extensive empirical study on eight datasets across three common IE tasks: Named Entity Recognition (NER), Relation Extraction (RE), and Event Detection (ED). The results show that **1)** LLMs outperform SLMs only when the overall number of annotations is limited, *i.e.,* both label types[3] and the samples[4] per label are extremely scarce. However, **2)** when we increase the number of samples (*e.g.,* a few hundreds), SLMs outperform LLMs by a large margin. We speculate it is caused by the natural limitations of ICL in two main ways. First, only a small subset of available samples can be used as demonstrations (demos) to prompt LLMs due to the maximum input length

[1] All LLMs discussed in this paper are not fine-tuned, and results for LLMs are based on ICL through APIs.

[2] We define SLMs as relatively small pre-trained language models (PLMs) that can be easily fine-tuned locally, such as BERT, RoBERTa, BART and T5.

[3] Label types denote *entity/relations/event types* in different tasks. We use them interchangeably there-in-after.

[4] Samples refer to (i) demonstrations in ICL of LLMs, or (ii) training samples for SLMs' fine-tuning.

of ICL. Moreover, more samples in demos might not bring extra performance gains. Second, as the schema (or number of label types) grows, the number of samples per label in prompts decreases. It is thus difficult to well understand tens (even hundreds) of label types and their semantic interactions via instruction. Besides, **3)** calling LLMs API suffers from much higher inference latency and financial cost than finetuning SLMs locally, especially when there are excessively long demos in ICL. Therefore, we claim that **LLM is not a good few-shot information extractor in general**.

We next investigate whether LLMs and SLMs exhibit different abilities to handle various types of samples, hoping LLMs could complement SLMs. We partition the testing samples into different groups according to their *difficulty* (measured by the confidence score of SLMs-based models) and compare the results of LLMs and SLMs on each group. We find that **4) LLMs are good at hard samples, though bad at easy samples**. We speculate the *hard* samples (*i.e.,* low confidence scores) require external knowledge or complex reasoning, which go beyond the abilities of SLMs but could be well solved by LLMs. For relatively *easy* samples, however, SLMs learn them well by fine-tuning parameters and perform much better than LLMs.

Building on these findings, we propose a novel adaptive *filter-then-rerank* framework to combine SLMs and LLMs considering both performance and cost in practice. The basic idea is that SLMs serve as a filter and LLMs as a reranker. In specific, SLMs make the first round of prediction, and if the sample is a hard one, we further pass the top-N candidate labels with highest prediction scores by SLMs to LLMs for reranking. The reranking mechanism leverages both LLMs and SLMs to deal with samples on which they are more proficient and thus combines their strengths. Moreover, it reranks only a small subset of test samples and minimizes the extra latency and budgets for calling LLM APIs. We outline our main contributions as follows:

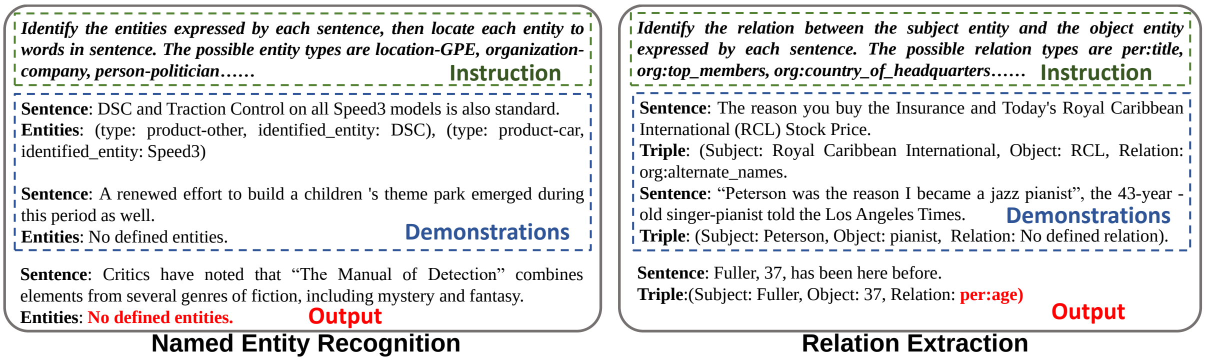
- We conduct an extensive empirical study comparing LLMs and SLMs on IE tasks. Our findings suggest that LLMs are generally not well-suited for IE tasks, especially when dealing with many samples and complicated schema.
- We propose a filter-then-rerank paradigm that combines the strengths of both LLMs and SLMs. We note that LLMs can be effective rerankers for challenging samples.
- With only 0.5%-13.2% of the samples being reranked, our adaptive filter-then-rerank system surpasses the previous state-of-the-art methods by an average 2.1% F1-score gain.

2 Related Work

Large Language Models (LLMs) We are fortunate to witness the emergent abilities (Wei et al., 2022b) of Large Language Models (LLMs, Brown et al. 2020; Chowdhery et al. 2022; Hoffmann et al. 2022) very recently. With qualitative progress on model scales (parameters, training corpus, training compute, etc. Brown et al. 2020; Chowdhery et al. 2022) and training strategies (code tuning, instruction tuning, human feedback, chain-of-thought tuning, etc. Chen et al. 2021; Wei et al. 2022a; Ouyang et al. 2022; Chung et al. 2022), LLMs show unprecedented reasoning and/or memorization abilities and benefit diverse NLP tasks.

In-context Learning (ICL) In our work, we use LLMs via ICL since fine-tuning LLMs for every downstream task is not practical. ICL enables LLMs to learn tasks through instructions and/or a few exemplars at inference stage without the need for model parameter updating. There are various approaches to improve the ICL ability of LLMs: (1) *Chain-of-Thought Reasoning (CoT)* (Wei et al., 2022c; Kojima et al., 2022; Zhang et al., 2023) leverages manual or auto-generated rationales to elicit the power of LLMs. (2) *Demonstration Selection (DS)* (Liu et al., 2022; Rubin et al., 2022; Su et al., 2022) retrieves appropriate samples as demos via unsupervised sentence similarity or supervised neural retriever. (3) *Self-consistency* (Wang et al., 2023) or *Self-ensemble (SE)* runs LLMs multiple times and determines the final results from different results by majority voting. We explore all three variants in Section 3.3.

ICL in Information Extraction Tasks There have been two branches exploring the use of LLMs in few-shot IE tasks with the aid of ICL. The first branch (Ding et al., 2022; Josifoski et al., 2023) views LLMs as an annotator and generates abundant samples with (pseudo) labels via ICL approaches. They then train SLMs using augmented data to achieve superior performance. Another branch, which includes our work, directly employs LLMs for inference. We contend that our approach differs from theirs in at least two ways. Firstly, previous work has been constrained to a single task type or a limited number of label types, and



Identify the entities expressed by each sentence, then locate each entity to words in sentence. The possible entity types are location-GPE, organization-company, person-politician……

Instruction

Sentence: DSC and Traction Control on all Speed3 models is also standard.
Entities: (type: product-other, identified_entity: DSC), (type: product-car, identified_entity: Speed3)

Sentence: A renewed effort to build a children 's theme park emerged during this period as well.
Entities: No defined entities.

Sentence: Critics have noted that "The Manual of Detection" combines elements from several genres of fiction, including mystery and fantasy.
Entities: No defined entities.

Identify the relation between the subject entity and the object entity expressed by each sentence. The possible relation types are per:title, org:top_members, org:country_of_headquarters……

Sentence: The reason you buy the Insurance and Today's Royal Caribbean International (RCL) Stock Price.
Triple: (Subject: Royal Caribbean International, Object: RCL, Relation: org:alternate_names.
Sentence: "Peterson was the reason I became a jazz pianist", the 43-year - old singer-pianist told the Los Angeles Times.
Triple: (Subject: Peterson, Object: pianist, Relation: No defined relation).

Sentence: Fuller, 37, has been here before.
Triple:(Subject: Fuller, Object: 37, Relation: per:age)

Figure 1: Prompts used in Vanilla ICL approach. We box out the instruction in green and demonstrations in blue. Actually the instruction lists all labels and demos usually contain tens of examples. Here we show only three labels and two examples for convenience of visualization. The outputs generated by LLMs are colored in red.

conducted on different experimental settings. For example, Wang et al. (2022) concentrated on the Event Argument Extraction (EAE) task, while Qin et al. (2023) focused solely on the NER task. Although Jimenez Gutierrez et al. (2022) studied both NER and RE tasks, they chose datasets with simple schemas (no more than 2 entity types and 5 relation types). These inconsistencies in experimental setups have led to inconclusive findings on whether LLMs outperform SLMs. To address the issue, our work concurrently tackles multiple tasks and conducts experiments on widely-used datasets with varying schema complexities ranging from 4 to 168 label types. Secondly, previous work solely relied on LLMs, while we have developed an adaptive *filter-then-rerank* approach based on comprehensive empirical research, which combines the strengths of both SLMs and LLMs.

3 Large LMs v.s. Small LMs

We wonder whether LLMs can outperform supervised SLMs in few-shot IE scenarios purely through ICL. To this end, we evaluate SLMs and LLMs on eight commonly used datasets spanning three representative IE tasks. (1) **Named Entity Recognition** (NER): CONLL'03 (Tjong Kim Sang and De Meulder, 2003), OntoNotes 5.0 (Weischedel et al., 2013) and FewNERD (Ding et al., 2021). (2) **Relation Extraction** (RE): TACRED (Zhang et al., 2017) and TACREV (Alt et al., 2020). (3) **Event Detection** (ED): ACE05 (Doddington et al., 2004), MAVEN (Wang et al., 2020) and ERE (Song et al., 2015). We list the statistics of these eight datasets in Appendix A.1.

3.1 Experimental Setup

We construct few-shot datasets from the original eight datasets mentioned above.

Training and Validation Set We adopt K-shot sampling strategy to construct few-shot datasets, *i.e.,* sampling K samples for each label type. We set 4 different K-values (1, 5, 10, 20) for NER and ED tasks, and 6 different K-values (1, 5, 10, 20, 50, 100) for RE tasks. For each constructed dataset with more than 300 sentences, we split 10% sentences as validation set and the remaining sentences as training set. See Appendix A.2 for details and the statistics of the sampled few-shot IE datasets.

Test Set To reduce the inference time and cost of LLMs, we randomly sample 250 sentences from the original test sets for NER and ED tasks, and 500 sentences for RE task, as our test benchmark.

Evaluation Metric We adopt micro-F1 score as evaluation metric. The reported value of each setting is the averaged score w.r.t 5 sampled train/validation sets to reduce random fluctuation.

3.2 Small Language Models

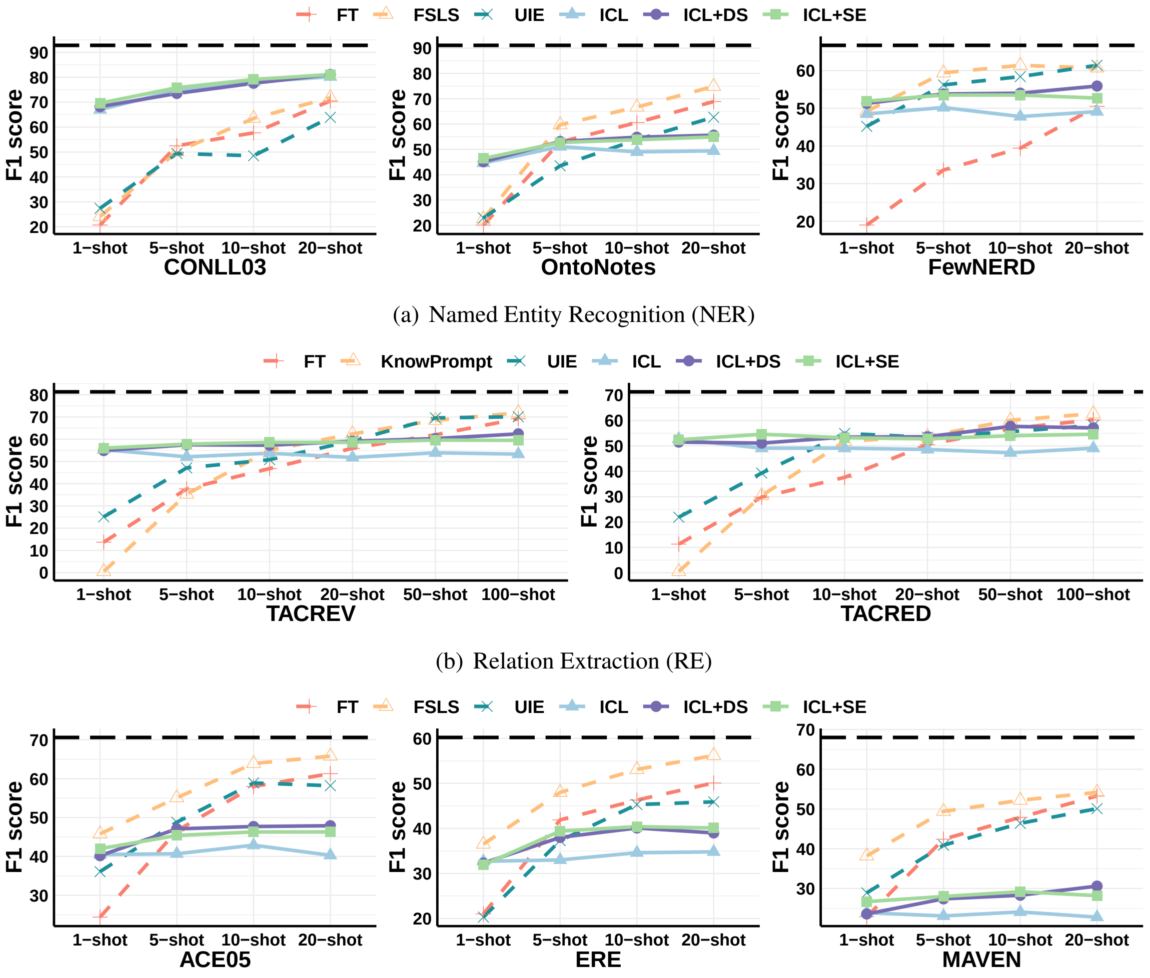
We adopt four representative supervised methods to evaluate the ability of SLMs on few-shot IE tasks. We choose `RoBERTa-large` (Liu et al., 2019) as the backbone of extractive-based methods and `T5-large` (Raffel et al., 2020) as the backbone of generation-based methods, respectively. Next, we brief these methods and leave their implementation details in Appendix B.1.

(1). Fine-tuning (FT): Add a classifier head on SLMs to predict the labels of each sentence/word.

(2). FSLS (Ma et al., 2022): The state-of-the-art extractive-based method for few-shot NER task. Anonymous (2022) also validate its competitive performance on few-shot ED tasks.

(3). KnowPrompt (Chen et al., 2022): The best extractive-based method for few-shot RE task.

(4). UIE (Lu et al., 2022b): A competitive unified generation-based method for few-shot IE tasks.



(a) Named Entity Recognition (NER)

(b) Relation Extraction (RE)

(c) Event Detection (ED)

Figure 2: Overall results of 4 supervised SLM-based methods (dashed lines) and 3 LLM-based ICL methods (solid lines) on eight datasets. **FT**: Fine-tuning. **ICL**: In-context Learning. **DS**: Demonstration Selection. **SE**: Self-ensemble. The results are averaged over 5 repeated experiments. See detailed values in Tables 11, 12 and 13.

3.3 Large Language Models

We adopt the current[5] strongest **CODEX**[6] (Chen et al., 2021) rather than **InstructGPT**[7] (Ouyang et al., 2022) in our experiments out of two primary reasons: (1) They share similar scales (~175B) and functionality. Previous study (Wang et al., 2022) showed CODEX has comparable, if not better, ICL capability. Our pivot experiments in Appendix C.1 further validates it. (2) Calling InstructGPT via API[8] is very costly.[9] In contrast, CODEX is currently available for free, allowing us to reduce cost and improve the reproducibility of our experiments.

We adopt several ICL approaches to evaluate the LLM ability on IE tasks. We introduce them in brief here and leave their details in Appendix B.2.

(1). Vanilla ICL utilizes the common prompts consisting of instruction, demonstrations (demos) and question. We show such format in Figure 1.

(2). ICL w. Automatic Chain-of-thought (Auto-CoT, Zhang et al. 2023) first bootstrap rationales from original examples. These generated rationales then act as intermediate reasoning steps in demos.

(3). ICL w. Demonstration Selection (DS, Liu et al. 2022) retrieves similar training examples as demos for each test example. We adopt a heuristic unsupervised approach here measuring the similarity of each sentence by their embeddings.

(4). ICL w. Self-ensemble (SE) predicts each test example by multiple times. Each time the examples (and/or their orders) are different in the demos. These predicted outputs are ensemebled by major-voting to determine the final prediction. [10]

[5] `2023-03-03`

[6] `code-davinci-002`

[7] `text-davinci-003`

[8] https://openai.com/

[9] The estimated cost of using InstructGPT is about 40K dollars to reproduce all the experiments in this section.

[10] Note that this approach is different from ICL with self-

Table 1: The inference seconds over 500 sentences. **B**: `RoBERTa/T5-base`. **L**: `RoBERTa/T5-large`. **C**: `code-davinci-002`. **T**: `text-davinci-003`.

	Task	**FT** (Roberta)		**UIE** (T5)		**ICL** (GPT-3)	
		B	L	B	L	C	T
NER	CONLL03	0.6	1.6	3.0	10.3	128.8	113.8
	OntoNotes	1.6	4.1	9.3	22.9	134.1	114.6
	FewNERD	1.1	2.8	14.6	39.4	179.4	166.5
RE	TACRED	0.4	1.4	14.1	43.8	164.4	132.4
	TACREV	0.4	1.4	14.5	45.6	151.6	127.1
ED	ACE05	0.8	2.4	3.0	8.9	135.2	112.1
	ERE	0.9	1.9	5.2	15.8	136.6	102.2
	MAVEN	2.6	6.6	31.5	62.5	171.7	156.2

3.4 Comparison Result

We evaluate eight approaches, four SLM-based supervised methods and four LLM-based ICL methods introduced above, on eight datasets across three IE tasks. We first conduct pivot experiments and observe ICL with Auto-CoT delivers much poorer results than we expected (see results and analysis in Appendix C.2). Therefore we do not include ICL with Auto-CoT in main experiments.

We first overview the performance of the remaining seven methods in Figure 2.

Comparison among ICL Approaches We observe that vanilla ICL achieves the worst results among three remaining ICL approaches. Since the number of examples in demos is bounded by LLMs' maximum input length, the increase of sample number brings no benefit and the performance of ICL is at a standstill once the sample number exceeds some threshold. Both DS and SE slightly alleviate such problem to some extent. Overall speaking, SE is better than DS with fewer shots, while DS outperforms the other two methods with more samples, *i.e.,* more retrieval candidates.

LLMs are not good few-shot Information Extractor. Even compared with the best LLM-based methods, SLMs mostly outperform by a large margin . (1) For most of the NER and RE datasets (except CONLL'03 with only 4 defined entities), the best LLM-based methods merely present significantly superior performance than SLMs under extremely low-resource settings (1-shot). The most competitive SLMs usually achieve comparable results with LLMs under 5-10 shot settings. With more samples, LLM-based methods perform worse than almost all SLM-based methods. (2) For three ED datasets, SLMs consistently beat all LLM-based methods even under 1-shot settings.

LLMs show limited inference speed We additionally compare the inference speed of different methods and show their results in Table 1. Expectedly we observe that the most efficient LLM-based method, *i.e.,* vanilla ICL, is still much slower than SLMs since they have much more parameters and longer input context length.

3.5 Discussion: Why LLMs not present satisfactory performance on IE tasks?

We dive deep into above results and analyze why LLMs fail to achieve satisfactory performance.

Underutilized Annotations We notice SLMs benefit much more than LLMs from extra annotations, *i.e.,* more training samples and label types. We speculate LLMs are constrained by ICL from two aspects. **(1) More samples**: The number of actual effective samples for LLMs, *i.e.,* the sample number in demos, is bounded by the maximum input length. Furthermore, we observe that in some tasks, the performance of LLMs have been at a standstill before reaching the input bounds (see Appendix C.4). In contrast, SLMs could learn from much more samples by parameter updating. Therefore the average difference between SLMs and LLMs grows as the K-shot value increases. **(2) More labels**: LLMs perform relatively best on CONLL'03 dataset with only 4 defined entities, and perform worst on MAVEN dataset with 168 event types. As Valmeekam et al. (2022) suggest, too many labels are likely to cause LLMs hard to understand all labels and their semantic interactions from the provided instruction and demos. Moreover, the number of examples per label in demos decreases as the number of label types increases.

Unexplored Task format We find LLMs achieve relatively worse performance on NER and ED tasks. We speculate it is partly due to their task formats. Both of these two tasks require structured outputs, *i.e.,* the (label, span) tuples as shown in Figure 1. Moreover, the number of outputs and the extracted span within each output are not fixed. Standing with Josifoski et al. (2023), we believe ICL approaches are not experienced on such task formats.

consistency (Wang et al., 2023). The randomness in self-consistency lies in output rationales (and answers) generated from nucleus sampling (Holtzman et al., 2020). The randomness in self-ensemble, however, comes from the different examples in input demos. We find setting the samping temperature coefficient $t = 0$, *i.e.,* greedy decoding, achieves the optimal result according to pivot experiments in Appendix C.3. Therefore we use **self-ensemble** rather than **self-consistency** (which requires a non-zero t) in this work.

Abstract Event Understanding We observe that ED achieves the worst performance among three tasks. In addition to its task format, (1) the definition of events is more abstract than that of entity and relations, (2) and the rules for labeling events are more complicated[11]. Therfore we speculate the abilities required to solve ED task are unlikely to be learned during the pre-training of LLMs or be generalized through instructions during ICL.

4 LLMs are Good Few-shot Reranker

Despite their unsatisfactory performance and their significant time and monetary costs, we believe that LLMs' abilities in memorization and reasoning remain crucial strengths for solving IE tasks. Consequently, we aim to identify more effective methods to harness the strengths of LLMs while minimizing their limitations in this section.

We propose a novel *filter-then-rerank* paradigm. Based on such paradigm, we observe the complementary results of LLMs and SLMs on samples with varying levels of difficulty. We would detail them in the following two subsections.

4.1 Filter-then-rerank Paradigm

Our *filter-then-rerank* paradigm combines SLMs and LLMs, as its name implies, by utilizing both as a filter and a reranker within it. SLMs act as a filter which eliminates unlikely labels and retains only the top-N candidates. LLMs then rerank these N labels and output the final answer.

Within our *filter-then-rerank* paradigm, we dynamically obtain N candidate labels for each test sample to be reranked by LLMs. We argue the prompts used in Section 3.3, which contain the whole schema in their instructions, are no longer needed. Instead, we reformulate the reranking procedure as a multiple-choice question. We show the format of our new multiple-choice question prompt as below, and leave real examples in Appendix F.

Instruction: Read the sentence and determine the relation between `<h ent>` and `<t ent>`.
Sentence: <Sentence>
(a) `T(<Label 1>, <h ent>, <t ent>)`
(b) `T(<Label 2>, <h ent>, <t ent>)`
(c) `T(<Label 3>, <h ent>, <t ent>)`
[Optional] Analysis: <Rationale>
Answer: <Correct choice>

As shown above, the instruction asks LLMs to determine the relation between entities in the sentence followed by. Then N candidate labels filtered from SLMs are provided. Each label is rephrased as a piece of choice text using the template `T(<label>, <h ent>, <t ent>)`. The template describes that the head entity `<h ent>` and the tail entity `<e ent>` have `<label>` relation. For example, `T`(cities_of_residence, Charles, Abidjan) = Charles lives in the city Abidjan. Then LLMs are expected to answer this multiple-choice question (optionally associated with rationales) as colored in red. We believe the format of multiple-choice question has various advantages: (1) It narrows the label scope significantly. (2) It lowers the difficulty of IE tasks since LLMs are more familiar with this format.

4.2 LLMs are and only are *Hard* Sample Solver

Unfortunately, we find such a *filter-then-rerank* system still performs far from satisfactory. Moreover, it leads to longer latency since LLMs rerank candidates per sample, rather than per sentence.

We intuitively speculate LLMs are more proficient than SLMs on *hard* samples. The *hard* samples here refer to those requiring external knowledge or complex reasoning beyond the capability of SLMs, see examples in Appendix E. SLMs could not solve them well with limited model capacity and data amount, while LLMs could solve them better. In contrast, SLMs could learn *easy* samples well from more samples by updating their parameters and exceed the performance of LLMs.

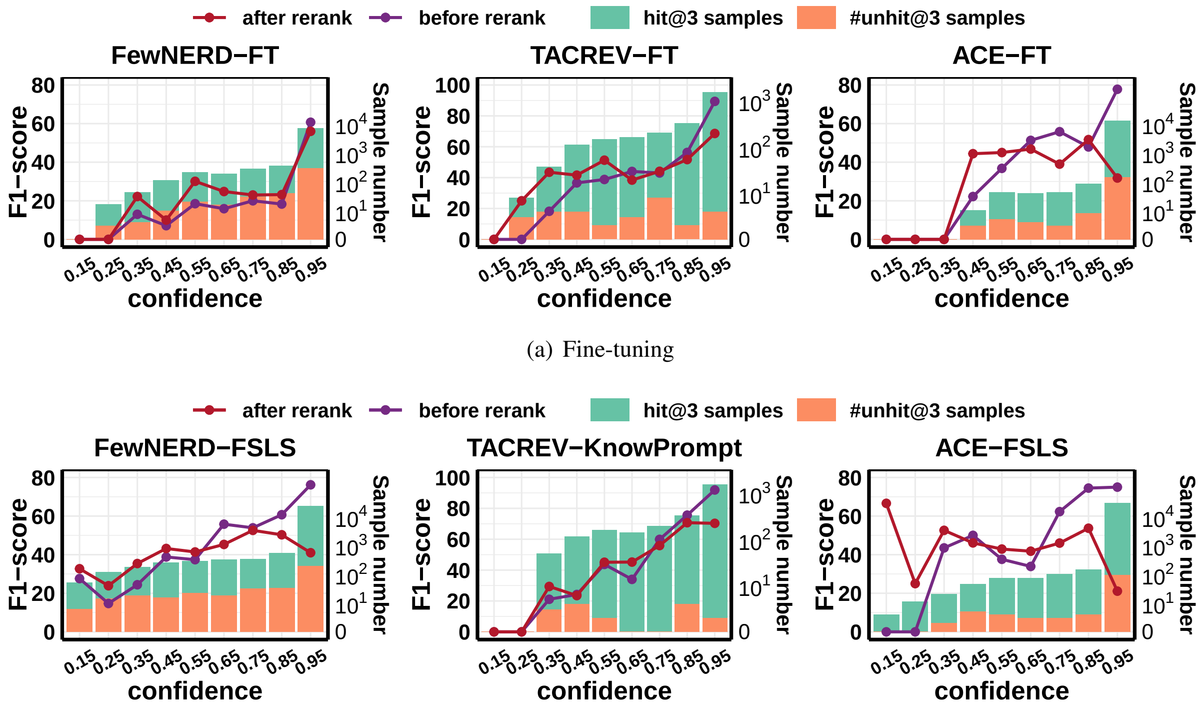
We design experiments to validate our conjecture. We group the samples by their *difficulty* and evaluate their performance before and after reranking within each group. To measure the *difficulty* of a sample x for SLMs, we adopt the maximum probability among all labels as the confidence score,

$$s(x) = \max_{y \in Y} p(y|x) \tag{1}$$

where $p(y|x)$ represents the probability of x having label y computed by SLMs. We call samples with low confidence scores as *hard* samples.

We conduct experiments on various datasets with SLM-based methods (*i.e.,* the filter). We select FewNERD, TACREV and ACE05 as the datasets, fine-tuning and the best baseline in Section 3.4 (FSLS for FewNERD and ACE05 datasets, KnowPrompt for TACREV dataset) as two SLM-based

[11] e.g., when a word triggers event and when does not, which word to annotate if more than one words trigger the event within the single sentence, and so on.



(a) Fine-tuning

(b) FSLS/KnowPrompt

Figure 3: Relationship between confidence score and F1-score of samples, before (colored in purple) and after (colored in red) LLM reranking. The x-axis shows 9 sample groups divided by their confidence scores. The left y-axis corresponds to the F1-score curves depicting the performance of samples within each group. The right y-axis (log scale) corresponds to the bars indicating the sample number in each group. Each subfigure is titled as `[X]-[Y]`, where `[X]` represents the dataset and `[Y]` represents the filter (SLM-based) methods. Among various datasets and the filter methods, we observe the similar trend between F1-scores before and after reranking.

methods. The results illustrated in Figure 3 validate our assumption. (1) There exists a consistent trend among different tasks and SLM-based methods that the performance of low-confidence samples improves after reranking while the high-confidence ones get degraded from reranking (see the red curves are usually higher than the purple curves when the confidence being less than a threshold, but lower after exceeding such threshold). In other word, **LLMs are more proficient than SLMs on hard samples, but often make mistake on easy samples**. (2) The hit@3 scores of SLMs usually achieve more than 95%[12] even under low-confidence scenarios. It ensures almost all true answers are included in candidate options for LLMs to rerank, making our reranking feasible. (3) The performance of LLMs might collapse under samples with very high-confidence. It is likely due to that LLMs sometimes tend to give false-positive predictions for negative samples, most of which are easy for SLMs and lie in high-confidence interval.

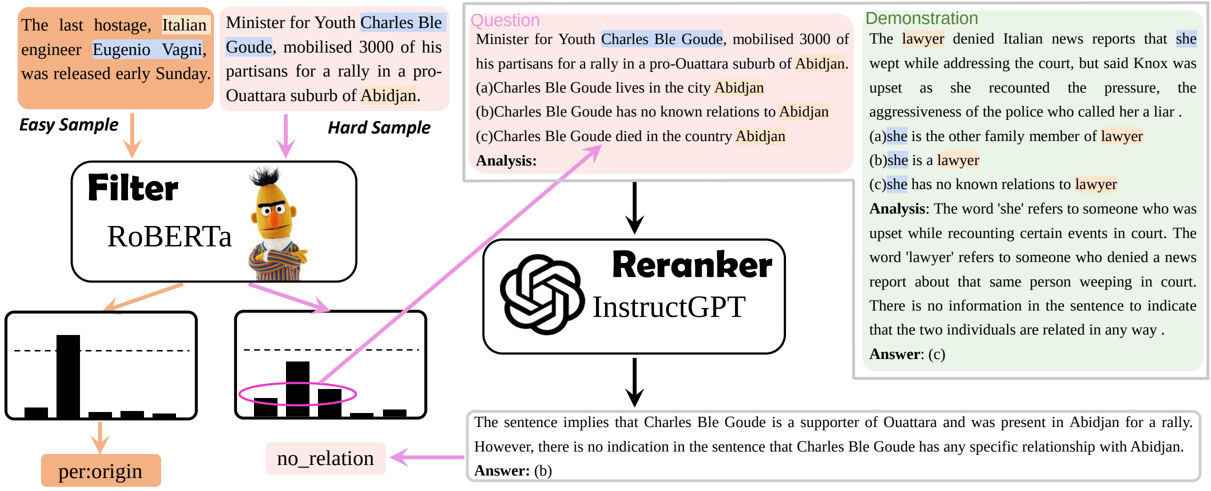
5 An Adaptive Reranker: Only Solve *Hard* Samples!

We summarize our findings above: (1) SLMs outperform LLMs under most scenarios, particularly with more training samples and complicated schema. (2) SLMs are much more lightwise and economical information extractor than LLMs. (3) LLMs could act as strong rerankers on *hard* samples with which SLMs fail to deal smoothly. Based on these findings, we propose a simple, efficient and effective adaptive reranker to incorporate the strength of SLMs and LLMs. With the minimal intervention of LLMs, *i.e.,* only reranking hard samples, our method shows consistent and significant improvement on three few-shot IE tasks, surpassing the SOTA by 2.1% absolute F1 gains on average.

5.1 Method

We call our method as *adaptive filter-then-rerank* and illustrate it in Figure 4. We first train SLMs with supervised approach and use it to predict each test sample. For samples with confidence score higher than a threshold, we retain their predictions from SLMs as final results. Otherwise we

[12] Note that the proportions of unhit samples, *i.e.,* the orange bar, are distorted Figure 3 due to the log scale.



The last hostage, Italian engineer Eugenio Vagni, was released early Sunday.

Minister for Youth Charles Ble Goude, mobilised 3000 of his partisans for a rally in a pro-Ouattara suburb of Abidjan.

Minister for Youth Charles Ble Goude, mobilised 3000 of his partisans for a rally in a pro-Ouattara suburb of Abidjan.
(a)Charles Ble Goude lives in the city Abidjan
(b)Charles Ble Goude has no known relations to Abidjan
(c)Charles Ble Goude died in the country Abidjan
Analysis:

The lawyer denied Italian news reports that she wept while addressing the court, but said Knox was upset as she recounted the pressure, the aggressiveness of the police who called her a liar .
(a)she is the other family member of lawyer
(b)she is a lawyer
(c)she has no known relations to lawyer
Analysis: The word 'she' refers to someone who was upset while recounting certain events in court. The word 'lawyer' refers to someone who denied a news report about that same person weeping in court. There is no information in the sentence to indicate that the two individuals are related in any way .
Answer: (c)

The sentence implies that Charles Ble Goude is a supporter of Ouattara and was present in Abidjan for a rally. However, there is no indication in the sentence that Charles Ble Goude has any specific relationship with Abidjan.
Answer: (b)

Figure 4: The overall architecture of our adaptive *filter-then-rerank* paradigm. We color easy samples in orange and hard samples in pink. For easy samples, the final predictions are exactly from the SLM-based methods. For hard samples, the top-N predictions from SLMs are fed into LLMs as the format of multiple-choice questions (pink box). The question is paired with demos (green box, we only provide 1 demo for convenience of visualization). LLMs rerank these N candidates and generate the final prediction.

select their top-N predictions and rerank them via LLMs. Here the threshold is adaptively determined by maximizing F1-score of the validation set.

As shown in right part of Figure 4, the prompt used in our adaptive reranker is composed of demos and an unanswered question. The demos (green part) contain several exemplars and each of them is an answered multiple-choice question as shown in Section 4.1. The unanswered question (pink part) is exactly a *hard* sample with N candidate labels to be reranked. The LLMs would rerank the candidates by answering this question.

5.2 Experimental Setup

We conduct experiments on FewNERD for NER task, TACREV for RE task and ACE05 for ED task. We set K as (5, 10, 20) for NER/ED tasks, and (20, 50, 100) for RE task. We use the competitive SLM-based methods shown in previous experiments (FSLS for NER and ED tasks, KnowPrompt for RE task. Both use `RoBERTa-large` as backbones) as the filter, and a 175B InstructGPT (`text-davinci-003`) as the reranker.

We select the top-3 candidate labels for NER/RE tasks, and top-2 candidate labels for ED task from SLMs' predictions and feed them to LLMs. We also add `No-label` choice if it is not in the top-N predictions of each sample. We select 20 samples from validation set and write their rationales manually for each dataset. We randomly choose 4 example from them as demos each time. See some demo examples in Appendix F. We follow template in Lu et al. (2022a) as our choice-template for TACREV dataset, and write templates used in FewNERD and ACE05 dataset by ourselves. We list all these templates in Appendix G.

5.3 Main Results

We compare 5 methods in this section to validate the effectiveness of our *filter-then-rerank* method.

(1) LLM-based ICL approach: We select the most competitive LLM-based variant shown in Section 3.4, ICL with DS, as a baseline.

(2) FSLS/KnowPrompt: The most competitive SLM-based methods (FSLS for NER and ED tasks, KnowPrompt for RE task) shown in Section 3.4.

(3) FSLS/KnowPrompt + Ensemble: We ensemble two SLMs (trained with the same dataset but different seeds), to validate the score gains from reranking is not only due to the ensemble effect.

(4) FSLS/KnowPrompt + Rerank: Our adaptive reranker with **one** SLM model as the filter.

(5) FSLS/KnowPrompt + Ensemble+ Rerank: Our adaptive reranker with ensemble of **two** SLMs as the filter, to further validate the gains from ensemble and reranking are complementary.

We overview the results listed in Table 2 and observe that our *filter-then-rerank* method achieves consistent and significant improvement on nine different settings across three datasets. The reranker brings a 2.4% average F1-gains without ensemble (line 3 v.s 5) and 2.1% F1-gains with ensemble (line 4 v.s 6). Thus we conclude that (1) the reranking approach is effective, and (2) the gains it brings are different and complementary to the ensemble.

Table 2: Overall results of LLM-based ICL methods, SLM-based supervised methods, and our proposed *filter-then-rerank* (S+L) methods. The best results are in bold face and the second best are underlined. All results except InstructGPT are averaged over 5 repeated experiments, and sample standard deviations are in the round bracket (the same below). The standard deviations are derived from different sampling few-shot datasets instead of random seeds. Thus high standard deviation values do not mean that no significant difference among these methods.

	Method	FewNERD			TACREV			ACE		
		5-shot	10-shot	20-shot	20-shot	50-shot	100-shot	5-shot	10-shot	20-shot
LLM	CODEX	53.8 (0.5)	54.0 (1.4)	55.9 (0.5)	59.1 (1.4)	60.3 (2.4)	62.4 (2.6)	47.1 (1.2)	47.7 (2.8)	47.9 (0.5)
	InstructGPT	53.6 (–)	54.6 (–)	57.2 (–)	60.1 (–)	58.3 (–)	62.7 (–)	52.9 (–)	52.1 (–)	49.3 (–)
SLM	FSLS / KnowPrompt	59.4 (1.5)	61.4 (0.8)	60.7 (1.9)	62.4 (3.8)	68.5 (1.6)	72.6 (1.5)	55.1 (4.6)	63.9 (0.8)	65.8 (2.0)
	+ Ensemble	59.6 (1.7)	61.8 (1.2)	62.6 (1.0)	64.9 (1.5)	71.9 (2.2)	74.1 (1.7)	56.9 (4.7)	64.2 (2.1)	66.5 (1.7)
S+L	+ LLM Rerank	60.6 (2.1)	62.7 (0.8)	63.3 (0.6)	66.8 (2.6)	72.3 (1.4)	75.4 (1.5)	57.8 (4.6)	**65.3** (1.7)	67.3 (2.2)
	+ Ensemble + LLM Rerank	**61.3** (1.9)	**63.2** (0.9)	**63.7** (1.8)	**68.9** (1.3)	**74.8** (1.3)	**76.8** (1.2)	**59.5** (3.7)	65.3 (1.9)	**67.8** (2.1)

Table 3: Ablation study on three datasets. 20-shot settings for FewNERD and ACE05. 100-shot setting for TACREV. The filter is two ensembled SLMs.

Manual CoT	Demo Selection	Cand Filter	FewNERD	TACREV	ACE05
✓	✓	✓	**63.7** (1.8)	**76.8** (1.2)	**67.8** (2.1)
✗	✓	✓	63.4 (1.2)	74.9 (1.7)	66.7 (2.2)
✗	✗	✓	63.1 (1.3)	74.6 (2.7)	66.8 (2.5)
✗	✓	✗	62.7 (1.3)	73.9 (1.1)	66.9 (2.4)
✗	✗	✗	62.7 (0.9)	72.8 (3.2)	66.2 (1.7)
FSLS/KnowPrompt			62.6 (1.0)	74.1 (1.7)	66.5 (1.7)

Table 4: The F1-score differences of all samples (the left three columns), reranked samples (the middle three columns), and the ratio of reranked samples (the last column) over three datasets. 20-shot settings for FewNERD and ACE05. 100-shot setting for TACREV. The filter is two ensembled SLMs.

	Overall			Reranked			Reranked Ratio
	before	after	△	before	after	△	
FewNERD	62.6	63.7	1.1	31.4	28.3	−3.1	3.3%
TACREV	74.1	76.8	2.7	33.8	43.4	9.6	13.2%
ACE05	66.5	67.8	1.3	35.6	55.7	20.1	0.5%

5.4 Ablation Study

We investigate the effectiveness of modules in LLM-reranker by removing each of them in turn. (1) **Manual CoT**: We remove the rationales of examples in demos. (2) **Demonstrations**: We further remove all demos, making the reranking procedure as a zero-shot QA (question answering) problem. (3) **Candidate Filter**: We keep all samples rather than only top-N labels for reranking.

We show their results in Table 3 and observe that (1) Demos with manual CoT achieves consistent improvement on all three datasets. It indicates that the rationales on correct choices further elicit the reranking ability of LLMs. (2) Even the demos without rationales improve the performance to some extent, see the comparison line 2 v.s. 3 and line 4 v.s. 5. (3) The filtering of candidate labels usually brings gains, especially on TACREV dataset. Furthermore, it significantly reduces the input length of LLMs and thus the inference cost.

5.5 Analysis

Few makes big difference We know from Figure 3 that most of samples are *easy* for SLMs (with high confidence score). Therefore only a tiny minority of samples are fed to LLMs for reranking, as shown in Table 4 (the last column). However, we figure out from Table 4 that the reranking brings impressive improvement on them, see ~10% for TACREV and ~20% for ACE05 dataset. Such upheaval on small amount of samples leads to an overall significant improvement.[13]

Few makes small cost We compare the inference cost between direct ICL via InstructGPT and *filter-then-rerank* method from two aspects, financial and time, in Figure 5. It shows that *filter-then-rerank* achieves a reduction of ~80%-90% on budgets and latency compared with the direct ICL methods. We point out that it is much more efficient due to three main reasons: (1) It calls LLMs API for only a small portion of samples. (2) It reranks only a small set of labels. (3) It requires less demos.

[13] An exception occurs at FewNERD dataset, on which the performance of reranked samples seems to degrade slightly. We dive deep into the result and observe that LLM-rerankers correct (eliminate, in other word) many false-positive samples. Therefore the overall performance actually improves even though the metric values on reranked samples decrease.

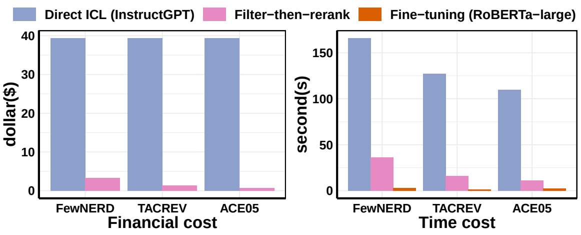


Figure 5: The financial and time cost over 500 sentences among three kinds of methods. Note that we do not take the financial cost of SLMs into account since they could run locally, and the financial cost of LLMs are estimated.

6 Conclusion

We have conducted an extensive empirical study comparing LLMs and SLMs on eight datasets across three tasks. We show that LLMs are still not good few-shot information extractor due to the task format, limited sample capacity and oversized schema. Meanwhile, compared to SLMs, LLMs incur significant time and monetary costs. We discover, however, LLMs could largely help SLMs to rerank and correct *hard* samples. Building on these findings, we propose an adaptive "filter-then-rerank" paradigm that leverages the strengths of both LLMs and SLMs while avoiding their limitations. This approach consistently yields promising results, with 2.1% F1-gain on average on several few-shot IE tasks, while minimizing the cost of latency and budgets caused by calling LLM APIs.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Anonymous. 2022. Few-shot event detection: An empirical study and a unified view. *ACL Rolling Review (October)*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2778–2788. ACM.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator?

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, abs/2205.11916.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2022a. Summarization as indirect supervision for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6575–6594, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022b. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022. Label semantics for few shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1956–1971, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan

Lowe. 2022. Training language models to follow instructions with human feedback.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver?

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2022. Selective annotation makes language models better few-shot learners.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In *International Conference on Learning Representations*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change).

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Xingyao Wang, Sha Li, and Heng Ji. 2022. Code4struct: Code generation for few-shot structured prediction from natural language.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, Quoc Le, and Denny Zhou. 2022c. Chain of thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375, Online. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *International Conference for Learning Representation (ICLR 2023)*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

A Datasets

A.1 Full Datasets

We construct few-shot IE datasets and conduct an empirical study on eight datasets spanning three tasks, **with varying schema complexities**. We show their statistics in Table 5.

A.2 Details of Few-shot IE Datasets

Sampling Algorithm for Train/Valid Datasets. We downsample sentences from original training dataset to construct few-shot training and valid datasets. We adopt K-shot sampling strategy that each label has (at least) K samples. For RE task, each sentence has exactly one relation and we simply select K sentences for each label. For NER and ED tasks, each sentences is possible to contain more than one entities/events. Since our sampling is at sentence-level, the algorithm of accurate sampling , *i.e.,* finding exactly K samples for each label, is NP-complete[14] and unlikely to find a practical solution. Therefore we follow Yang and Katiyar (2020); Ma et al. (2022) adopting a greedy sampling algorithm to select sentences for NER and ED tasks, as shown in Algorithm 1. Note that the actual sample number of each label can be larger than K under this sampling strategy. For all three tasks, we additionally sample negative sentences (without any defined labels) and make the ratio of positive sentences (with at least one label) and negative sentences as 1:1. The statistics of the curated datasets are listed in Table 6.

Based on the subsets constructed above, we optionally further split them into training and valid sets. For few-shot datasets with more than 300 sentences, we additionally split 10% sentences as the valid set and the remaining sentences as training set. Otherwise, we do not construct valid set and conduct 5-fold cross validation to avoid overfitting.

Algorithm 1 Greedy Sampling

```
Require: shot number $K$, original full dataset $\mathcal{D} = \{(\mathbf{X}, \mathbf{Y})\}$ tagged with label set $E$
Sort $E$ based on their frequencies in $\{\mathbf{Y}\}$ as an ascending order
$\mathcal{S} \leftarrow \phi$, Counter $\leftarrow$ dict()
for $y \in E$ do
    Counter$(y) \leftarrow 0$
end for
for $y \in E$ do
    while Counter$(y) < K$ do
        Sample $(\mathbf{X}, \mathbf{Y}) \in \mathcal{D}$ s.t.$\exists j, y_j = y$
        $\mathcal{D} \leftarrow \mathcal{D} \backslash (\mathbf{X}, \mathbf{Y})$
        Update Counter (not only $y$ but all event types in $\mathbf{Y}$)
    end while
end for
for $s \in \mathcal{S}$ do
    $\mathcal{S} \leftarrow \mathcal{S} \backslash s$ and update Counter
    if $\exists y \in E$, s.t. Counter$(y) < K$ then
        $\mathcal{S} \leftarrow \mathcal{S} \bigcup s$
    end if
end for
return $\mathcal{S}$
```

B Models

B.1 Implementations Details on SLMs

We introduce the implementation details of 4 methods based on Small Language Models (SLMs).

Fine-tuning/FSLS. We implement these two methods by ourselves. We use `RoBERTa-base` and `RoBERTa-large` (Liu et al., 2019) as our backbones. We adopt Automatic Mixed Precision (AMP) training strategy[15] to save memory. We run each experiment on a single NVIDIA V100 GPU. We train each model with the AdamW (Loshchilov and Hutter, 2019) optimizer with linear scheduler and 0.1 warm-up steps. We set the weight-decay coefficient as 1e-5 and maximum graidient norms as 1.0. We set the batch size as 64, the maximum input sequence length as 192, the training step as 500 and the learning rate as 5e-5.

KnowPrompt We implement this method based on original source code[16], and use `RoBERTa-base` and `RoBERTa-large` as our backbones. We set 10 maximum epochs for 50- and 100-shot datasets, and as 50 epochs for other datasets. We keep all other hyperparameters as default, and run each experiment on a single NVIDIA V100 GPU.

UIE We implement this method based on original source code[17], and use `T5-base` and `T5-large` (Raffel et al., 2020) as the backbones. We run each experiment on a single NVIDIA Quadro RTX8000 GPU. We set the batch size as 16 with 1000 training steps for base model, and the batch size as 4 with 4000 training steps for large

[14] The *Subset Sum Problem*, a classical NP-complete problem, can be reduced to this sampling problem.

[15] https://pytorch.org/docs/stable/amp.html

[16] https://github.com/zjunlp/KnowPrompt

[17] https://github.com/universal-ie/UIE

Table 5: Statistics of three full ED datasets.

Dataset		Named Entity Recognition			Relation Extraction		Event Detection		
		CONLL	OntoNotes	FewNERD	TACREV	TACRED	ACE05	MAVEN	ERE
#Label Type		4	18	66	41	41	33	168	38
#Sents	Train	14,041	49,706	131,965	68,124	68,124	14,024	32,360	14,736
	Test	3,453	10,348	37,648	15,509	15,509	728	8,035	1,163
#Mentions	Train	23,499	128,738	340,247	13,012	13,012	5,349	77,993	6,208
	Test	5,648	12,586	96,902	3,123	3,123	424	18,904	551

Table 6: The statistics of few-shot training sets. We set different random seeds and generate 5 training sets for each setting. We report their average statistics.

Dataset Settings		# Labels	# Sent	# Sample	# Avg shot
CONLL'03	1-shot	4	4.8	5.8	1.4
	5-shot		16.2	21.8	5.5
	10-shot		29.2	42.6	10.7
	20-shot		65.6	82.0	20.5
OntoNotes	1-shot	18	20.0	33.4	1.9
	5-shot		84.8	148.0	8.2
	10-shot		158.6	281.0	15.6
	20-shot		332.8	547.2	30.4
FewNERD	1-shot	66	89.8	147.0	2.2
	5-shot		286.2	494.8	7.5
	10-shot		538.0	962.0	14.6
	20-shot		1027.2	1851.4	28.1
TACREV	1-shot	41	81.6	41.0	1.0
	5-shot		387.6	205.0	5.0
	10-shot		741.2	406.0	9.9
	20-shot		1367.2	806.0	19.7
	50-shot		2872.0	1944.0	47.4
	100-shot		4561.0	3520.0	85.9
TACRED	1-shot	41	81.6	41.0	1.0
	5-shot		387.6	205.0	5.0
	10-shot		741.2	406.0	9.9
	20-shot		1367.2	806.0	19.7
	50-shot		2871.2	1944.0	47.4
	100-shot		4575.2	3520.0	85.9
ACE05	1-shot	33	47.4	41.0	1.2
	5-shot		192.8	165.0	5.0
	10-shot		334.6	319.4	9.7
	20-shot		579.4	598.2	18.1
MAVEN	1-shot	168	157.6	298.0	1.8
	5-shot		540.4	1262.2	7.5
	10-shot		891.2	2413.8	14.4
	20-shot		1286.4	4611.4	27.4
ERE	1-shot	38	48.4	54.6	1.4
	5-shot		175.0	219.2	5.8
	10-shot		304.8	432.4	11.4
	20-shot		521.6	806.6	21.2

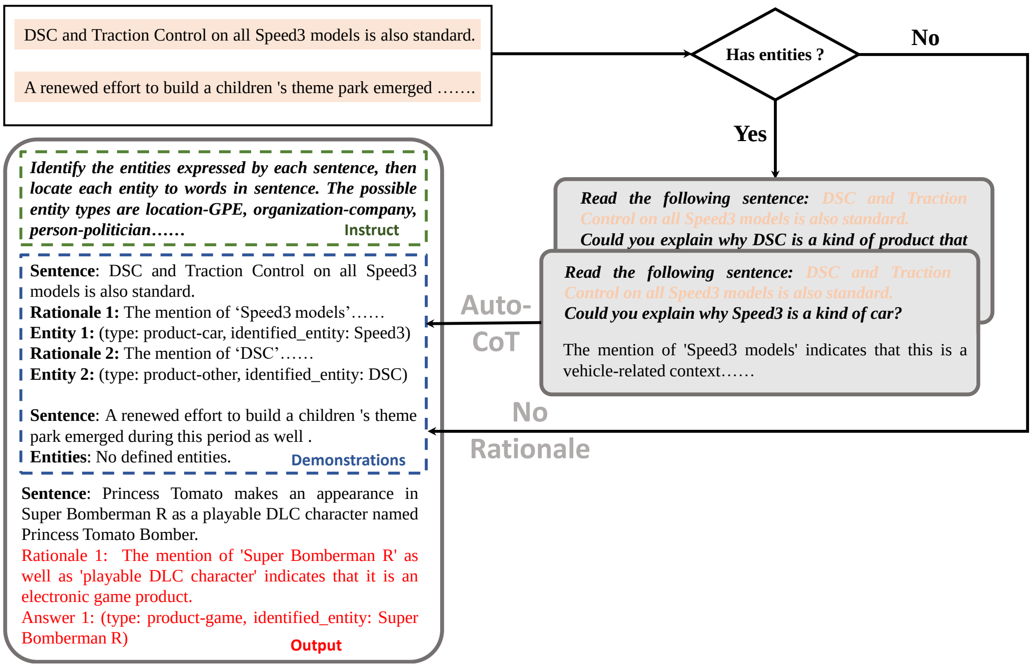
model. We set the maximum input sequence length as 800 and the learning rate as 1e-4.

B.2 Implementations Details on LLMs

We mainly use CODEX (Chen et al., 2021) as backbones of our ICL approaches. We also use InstructGPT (Ouyang et al., 2022) in pivot experiments (as shown in Appendix C) and our adaptive *filter-then-rerank* system. We set the maximum input length as 3600 for all tasks and models. The only exception occurs when we use CODEX to solve RE tasks: here we set the maximum input length as 7000. We unify the maximum output length as 96 for NER and ED tasks, and as 32 for RE task. We set the sampling temperature coefficient $t = 0$, *i.e.,* greedy decoding. We would detail the special design for each variant below.

Vanilla ICL Our prompts are composed of three parts: (1) `Instruction`: a short piece of natural language description about the task. (2) `Demonstrations`: some (input, output) pairs as train examples for LLMs. (3) `Question`: the input as test example. Most time the training sample number exceeds the limitation of examples in demos. Under this case, we would randomly sample a subset as demo examples for each test instance.

ICL w. Automatic Chain-of-thought (Auto-CoT, Zhang et al. 2023) If a sentence has positive labels, we would query LLMs *According to [sentence], Why [span] is a [label].* For example, given the sentence *"DSC and Traction Control on all Speed3 models is also standard."* in which *Speed3* is a *car-product* entity and *DSC* is an *other-product* entity, it is to auto-generate rationales for these two entities. By asking *"Could you explain why Speed3 is a kind of car"*, the LLMs would answer *"the term Speed3 refers to a specific car model produced by Mazda. Mazda is an automobile manufacturer, and as such, Speed3 is likely a car product from their lineup"* . As shown in Figure 6, we collect these auto-generated rationales


DSC and Traction Control on all Speed3 models is also standard.

A renewed effort to build a children 's theme park emerged …….

Identify the entities expressed by each sentence, then locate each entity to words in sentence. The possible entity types are location-GPE, organization-company, person-politician……

Sentence: DSC and Traction Control on all Speed3 models is also standard.
Rationale 1: The mention of 'Speed3 models'……
Entity 1: (type: product-car, identified_entity: Speed3)
Rationale 2: The mention of 'DSC'……
Entity 2: (type: product-other, identified_entity: DSC)

Sentence: A renewed effort to build a children 's theme park emerged during this period as well .
Entities: No defined entities.

Sentence: Princess Tomato makes an appearance in Super Bomberman R as a playable DLC character named Princess Tomato Bomber.
Rationale 1: The mention of 'Super Bomberman R' as well as 'playable DLC character' indicates that it is an electronic game product.
Answer 1: (type: product-game, identified_entity: Super Bomberman R)

Read the following sentence: *DSC and Traction Control on all Speed3 models is also standard.*
Could you explain why DSC is a kind of product that

Read the following sentence: *DSC and Traction Control on all Speed3 models is also standard.*
Could you explain why Speed3 is a kind of car?
The mention of 'Speed3 models' indicates that this is a vehicle-related context……

Figure 6: The diagram of ICL approach with Auto-CoT.

for each samples, and insert them between the sentences and answers (see blue box in left part of the figure). If a sentence has no positive labels, however, we do not ask LLMs and keep the original format as the vanilla ICL approach.

ICL w. Demonstration Selection (DS, Liu et al. 2022; Su et al. 2022) The format of prompts is the same as that of vanilla ICL approach. The difference from vanilla ICL lies in that we retrieve similar samples as demo examples for each instance rather than sample randomly. We adopt the cosine similarity of two sentence embeddings to measure their similarity. We compute sentence embeddings by SimCSE-RoBERTa$_{\texttt{large}}$ (Gao et al., 2021).

(4). ICL w. Self Ensemble (SE) We repeatedly test each sample 5 times with different demos and count the results. We select the winners with more than 1 votes as our final results.

C Pivot Experiments

LLMs require enormous financial and time cost during inference. Therefore we conduct several pivot experiments to prune some experimental settings with (1) potential unaffordable cost and (2) significant unsatisfactory performance.

C.1 CODEX v.s. InstructGPT

We tend to use CODEX rather than InstructGPT as much as possible in our empirical study since CODEX is now free access to the public. Therefore we choose one representative setting from each dataset and test the performance difference between CODEX and InstructGPT. We show their results in Table 8 and observe that there is no significant difference between these two LLMs. Based on this finding, we determine to only use CODEX for empirical study in Section 3.4.

Table 7: The F1-score difference between with and without Auto-CoT. We generate rationales by InstructGPT, then adopt **ICL w. Auto-CoT** approach and use CODEX as our backbone for inference.

10-shot train set	FewNERD	TACREV	ACE05
wo. Auto-CoT	**54.0**(1.4)	**57.3**(1.8)	**47.7**(2.8)
w. Auto-CoT	36.6(1.7)	22.0(1.2)	43.1(3.4)

We also find out, however, InstructGPT achieves much better results than CODEX in our adaptive *filter-then-rerank* system. Therefor we use InstructGPT in Section 4 and Section 5. Including this pivot experiments, we pay about 1000 dollars to call InsturctGPT API for this work.

C.2 ICL w. Auto-CoT

This section we explore whether ICL with Auto-CoT approach achieves competitive performance as we expected. Though CODEX achieves similar performance with InstructGPT on IE tasks, we do observe that InstructGPT is able to generate more fluent and reasonable explanations. Therefore we generate rationales using InstructGPT with tem-

Table 8: The F1-score difference between CODEX and InstructGPT. We adopt **ICL w. DS** approach for two LLMs.

	NER (20-shot)			RE (20-shot)		ED (20-shot)		
	CONLL	**OntoNotes**	**FewNERD**	**TACREV**	**TACRED**	**ACE05**	**MAVEN**	**ERE**
InstructGPT	77.2	47.7	**57.2**	**60.1**	52.1	**49.3**	**25.4**	**40.8**
CODEX	**81.1**	**55.6**	55.9	59.1	**53.6**	47.9	22.8	39.0

perature $t = 0.7$. We select several representative settings and compare the performance with and without Auto-CoT as shown in Table 7.

We are frustrated to find Auto-CoT degrades the performance with a large margin. We speculate this degration could be attributed to 3 main reasons. (1) The rationale increase the length of each sample and thus decrease the overall example number in demos. (2) There exists an obvious discrepancy between sentences with and without positive labels. As shown in Figure 6, the rationales are only provided for sentences with positive labels because it is hard to explain why a sentence dose not contain any label. (3) Some auto-generated rationales are low-quality, especially for RE tasks. We would explore better strategy to exploit auto-genertaed rationales in the future work.

C.3 Random Sampling for LLM Outputs

Previous work[18] tells us that it is better to set the sampling temperature $t = 0$ for tasks with structured outputs, including IE tasks. We validate this conclusion in Table 9, from which we could see the generated quality when $t = 0$ is much higher than the quality when $t \neq 0$. Therefore we set $t = 0$ in all main experiments, and do not take self-consistency (Wang et al., 2023) into account. Instead we adopt self-ensemble since it does not require the generation randomness.

Table 9: The F1-score difference between with and without non-zero t value.

10-shot train set	**FewNERD**	**TACREV**	**ACE05**
$t = 0$	48.5(1.9)	53.7(2.3)	42.9(2.2)
+ self-ensemble	**53.5**(1.3)	**58.6**(1.5)	**46.3**(0.8)
$t = 0.7$	40.9(2.3)	39.9(1.2)	35.6(1.0)
+ self-consistency	52.1(0.9)	53.4(1.3)	45.6(3.0)

C.4 Do More Samples in Demos help?

We wonder whether longer demos bring more performance gains. Thus we gradually increase the number of demos, and observe the changes of performance as the input length increases. We show the results in Figure 7 and observe that (1) The performance of RE task increase consistently. Thus RE task shows potential benefiting from more demos. (2) The performance of NER and ED gradually become stable even degrade with the increase of demos. It frustratingly implies these two tasks have been bounded even before achieving the maximum input length of LLMs.

D Terms

We summarize all terms used in our work and their abbreviations in Table 10.

Table 10: Term Table (with their abbreviations)

Abbreviation	Full Name
IE	Information Extraction
NER	Named Entity Recognition
RE	Relation Extraction
ED	Event Detection
LLMs	Large Language Models
SLMs	Small Pre-trained Langauge Models
ICL	In-context Learning
FT	Fine-tuning
Auto-CoT	Automatic Chain-of-thought
DS	Demonstration Selection
SE	Self-ensemble

E Case Study

Table 14 shows some *hard* examples which benefits from our adaptive *filter-then-rerank* paradigm. We could see that external knowledge (*e.g.,* Triptolemus is a figure in Greek mythology) and complex reasoning (*e.g.,* The coach of a Finland's football team may not be the residence of Finland) from LLMs do help to correct the errors made by SLMs.

F Demonstration examples

We convert few-shot IE tasks to multiple-choice questions in our *filter-then-rerank* paradigm. We show 4 examples used in demonstrations for FewNERD dataset in Table 15, for TACREV dataset in Table 16, and for ACE05 datasets in Table 17.

[18] https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api

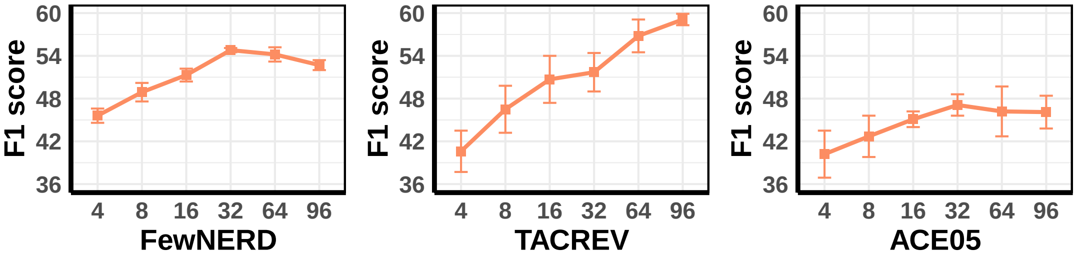


Figure 7: The F1-score difference with different demo number among three datasets: FewNERD for NER task, TACREV for RE task and ACE05 for ED task. We adopt **ICL w. DS** approach and use CODEX as the LLM in this experiment. The x-axis in each subfigure represents the number of demos (not the shot value K) during ICL.

Table 11: Performance with different methods on NER tasks. Averaged F1-scores with sample standard deviations on 5 repeated experiments are shown. The best results are in bold face and the second best are underlined.

	Method		FewNERD				OntoNotes				Conll			
			1-shot	5-shot	10-shot	20-shot	1-shot	5-shot	10-shot	20-shot	1-shot	5-shot	10-shot	20-shot
Small LMs	Fine-tune	Roberta-b (110M)	20.1(2.2)	33.7(3.4)	37.4(1.8)	48.6(1.5)	17.2(4.3)	50.2(3.1)	56.3(5.0)	64.5(3.6)	21.9(10.3)	50.1(7.0)	57.7(7.0)	64.8(4.9)
		Roberta-l (330M)	19.0(3.5)	33.6(3.5)	39.4(2.1)	50.5(1.4)	20.1(5.3)	52.9(3.3)	60.6(4.7)	68.9(4.8)	20.7(8.0)	52.5(8.8)	57.7(8.7)	70.5(2.4)
	FSLS	Roberta-b (110M)	46.5(1.3)	55.0(0.5)	57.3(1.7)	59.4(1.9)	29.0(4.5)	**60.8**(3.7)	65.1(5.7)	71.9(4.0)	28.6(9.0)	58.9(5.8)	63.9(4.8)	72.3(2.9)
		Roberta-l (330M)	49.2(2.8)	**59.4**(1.5)	**61.4**(0.8)	60.7(1.9)	22.5(4.3)	59.7(3.5)	**66.7**(3.6)	**74.9**(2.8)	24.1(10.9)	50.1(7.0)	63.5(3.5)	71.8(4.3)
	UIE	T5-b (330M)	41.5(3.3)	52.5(1.2)	55.4(2.3)	58.4(2.2)	21.3(4.1)	41.1(2.2)	48.7(4.3)	58.0(1.6)	23.8(7.9)	44.1(4.8)	54.5(6.9)	61.7(6.5)
		T5-l (770M)	45.2(2.8)	56.2(1.0)	57.7(3.1)	**61.4**(1.4)	22.9(3.1)	43.5(2.4)	53.8(2.5)	62.7(2.9)	25.0(7.8)	48.0(7.9)	50.1(9.8)	67.0(8.5)
Large LMs	ICL	CODEX (175B)	48.5(1.9)	50.2(1.0)	47.8(0.6)	49.1(1.7)	44.6(1.9)	51.1(1.8)	49.0(2.9)	49.4(1.8)	67.0(4.9)	74.4(4.3)	78.1(3.2)	80.2(2.1)
	ICL+DS	CODEX (175B)	51.3(1.9)	53.8(0.5)	54.0(1.4)	55.9(0.5)	45.1(1.6)	53.1(1.8)	54.8(1.5)	55.6(1.3)	68.2(4.9)	73.5(2.1)	77.6(2.5)	**81.1**(1.8)
	ICL+SE	CODEX (175B)	**51.9**(1.8)	53.5(0.6)	53.5(1.3)	52.7(0.8)	**46.4**(1.2)	52.7(1.4)	53.8(2.4)	54.9(1.0)	**69.6**(3.6)	**75.8**(3.3)	**79.1**(1.3)	81.1(2.4)

Table 12: Performance with different methods on RE tasks. Averaged F1-scores with sample standard deviations on 5 repeated experiments are shown. The best results are in bold face and the second best are underlined.

	Method		TARCREV						TARCRED					
			1-shot	5-shot	10-shot	20-shot	50-shot	100-shot	1-shot	5-shot	10-shot	20-shot	50-shot	100-shot
Small LMs	Fine-tune	Roberta-b (110M)	10.3(3.2)	35.0(3.1)	44.6(2.5)	54.1(1.3)	61.1(2.3)	66.4(2.1)	8.4(3.2)	25.7(4.1)	37.5(2.6)	47.2(1.0)	53.3(4.0)	59.7(1.4)
		Roberta-l (330M)	13.7(4.4)	37.7(3.5)	46.8(4.0)	55.9(4.4)	62.0(1.3)	69.2(3.7)	11.3(2.2)	29.8(4.6)	37.6(3.1)	50.5(3.6)	56.8(1.1)	60.3(4.6)
	KnowPrompt	Roberta-b (110M)	1.1(0.2)	33.1(6.6)	45.9(4.4)	57.8(2.8)	67.0(4.1)	68.1(1.6)	1.2(0.1)	25.0(4.9)	43.3(3.8)	49.3(2.5)	56.4(3.2)	60.1(1.8)
		Roberta-l (330M)	0.5(0.1)	35.3(4.8)	55.0(3.8)	**62.4**(3.8)	68.5(1.6)	**72.6**(1.5)	0.5(0.1)	30.4(6.2)	51.8(3.2)	**53.7**(3.6)	60.1(1.4)	62.7(2.0)
	UIE	T5-b (330M)	19.4(2.0)	44.1(1.4)	48.2(3.0)	51.9(3.5)	56.5(3.2)	59.3(2.7)	19.1(3.5)	34.3(2.0)	41.1(3.0)	49.4(1.5)	51.6(2.0)	53.1(1.4)
		T5-l (770M)	25.1(4.3)	47.1(2.2)	50.8(2.1)	59.3(1.9)	**69.6**(1.4)	70.1(1.2)	21.9(2.8)	39.3(1.9)	**54.9**(3.5)	53.4(3.2)	**61.5**(1.6)	**63.1**(2.5)
Large LMs	ICL	CODEX (175B)	55.3(1.8)	52.1(1.3)	53.7(2.3)	51.8(2.6)	53.9(1.4)	53.3(3.3)	**52.7**(1.7)	49.1(1.8)	49.1(1.5)	48.6(2.9)	47.3(2.0)	49.1(1.3)
	ICL+DS	CODEX (175B)	54.9(1.0)	57.5(3.6)	57.3(1.8)	59.1(1.4)	60.3(2.4)	62.4(2.6)	51.5(1.3)	51.1(3.0)	53.5(1.0)	53.6(1.3)	57.7(2.8)	57.1(1.3)
	ICL+SE	CODEX (175B)	**56.0**(1.1)	**57.8**(3.6)	**58.6**(1.9)	58.6(1.5)	59.6(0.3)	59.5(1.2)	52.4(2.1)	**54.6**(2.7)	53.2(0.7)	52.8(2.7)	54.0(1.3)	54.6(1.3)

Table 13: Performance with different methods on ED tasks. Averaged F1-scores with sample standard deviations on 5 repeated experiments are shown. The best results are in bold face and the second best are underlined.

	Method		ACE05				ERE				MAVEN			
			1-shot	5-shot	10-shot	20-shot	1-shot	5-shot	10-shot	20-shot	1-shot	5-shot	10-shot	20-shot
Small LMs	Fine-tune	Roberta-b (110M)	20.1(4.9)	45.1(6.6)	55.9(2.4)	60.3(1.9)	17.2(4.9)	35.5(2.1)	46.5(3.4)	47.3(2.5)	15.5(3.2)	36.1(1.9)	45.2(1.3)	52.4(1.4)
		Roberta-l (330M)	24.4(10.6)	46.6(4.5)	58.0(2.9)	61.3(1.6)	21.0(8.1)	41.9(4.8)	46.3(3.2)	50.1(2.8)	22.9(2.1)	42.4(2.4)	47.9(1.5)	53.3(0.6)
	FSLS	Roberta-b (110M)	34.8(6.6)	53.8(3.9)	58.6(3.4)	63.6(2.5)	27.9(5.8)	45.2(2.6)	51.6(1.3)	52.1(3.4)	35.7(2.3)	48.3(1.7)	50.5(1.9)	**54.7**(1.6)
		Roberta-l (330M)	**45.8**(6.2)	**55.1**(4.6)	**63.9**(0.8)	**65.8**(2.0)	**36.5**(3.0)	**48.0**(3.5)	**53.1**(2.8)	**56.2**(2.6)	**38.2**(2.4)	**49.4**(2.5)	**52.2**(0.8)	54.2(1.8)
	UIE	T5-b (330M)	32.2(4.4)	44.1(3.5)	48.3(4.5)	55.9(3.2)	20.9(4.1)	31.3(2.7)	37.0(3.8)	37.0(2.3)	23.9(1.7)	39.8(1.4)	43.8(1.4)	46.7(1.0)
		T5-l (770M)	34.7(3.3)	51.2(3.6)	58.9(3.3)	58.2(1.8)	20.3(4.1)	37.3(2.1)	45.3(2.4)	45.9(3.5)	28.9(3.0)	40.9(0.8)	46.4(1.7)	50.1(1.2)
Large LMs	ICL	CODEX (175B)	40.5(2.9)	40.7(2.0)	42.9(2.2)	40.3(2.9)	32.7(2.8)	33.0(1.6)	34.6(3.0)	34.8(2.5)	23.6(1.1)	27.4(2.0)	28.3(0.9)	30.6(0.9)
	ICL+DS	CODEX (175B)	40.2(2.3)	47.1(1.2)	47.7(2.8)	47.9(0.5)	32.3(3.6)	38.0(2.1)	40.1(1.3)	39.0(1.3)	23.9(1.9)	23.1(0.8)	24.1(1.0)	22.8(0.6)
	ICL+SE	CODEX (175B)	42.0(3.3)	45.4(1.5)	46.3(0.8)	46.3(1.3)	31.9(2.6)	39.4(4.2)	40.4(3.2)	40.1(2.4)	26.7(1.7)	28.0(0.3)	29.2(1.2)	28.2(1.0)

Table 14: Examples of the samples corrected by LLM reranking. We sample 3 examples from NER, RE and ED tasks, respectively. **Sentences**: The sentences in which samples locate. We color the samples (entities or trigger words) to be identified. **before**: The prediction before LLM reranking. Based on SLM-based methods. **after**: The reranked prediction using LLMs. **Rationales**: Auto-generated rationales by LLMs. Could be viewed as the explanation of the reranked results.

Sentences (entity/event)	Before	After	Rationales
According to it, King Eleusis and Cothonea (Cyntinia), parents of Triptolemus, are visited by Demeter, who rears their son, feeding him divine milk by day and placing him into the fire at night, which makes Triptolemus grow faster than mortal children normally do.	person-other ✗	other-god ✓	According to the context, Triptolemus is a figure in Greek mythology, the son of King Eleusis and Cothonea (Cyntinia). He was a demigod who was sent by Demeter to teach mankind the art of agriculture. So Triptolemus is a god in some legend/religious story.
Miettinen coaches Helsinki's PK-35, which shares the lead in Finland's top women's football league, and becomes the third Sky Blue coach in two months.	countries of residence ✗	None ✓	The sentence states that 'Miettinen coaches Helsinki's PK-35, which shares the lead in Finland's top women's football league.' This implies that Miettinen is a coach in Finland's top women's football league, but does not suggest any other specific relationship between Miettinen and Finland. So Miettinen has no known relations to Finland.
We have fought every institution, every law firm ... it 's cost us a fortune but now we 're going on a massive expansion.	Conflict.Attack ✗	None ✓	In the given sentence, the word fought is used to describe the action of the speaker fighting against various institutions and law firms. This does not involve any physical violence or court proceedings, so the word fought does not trigger any known event.

G Template

In our *filter-then-rerank* paradigm, we utilize templates converting candidate labels to question options. We list the template for FewNERD dataset in Table 18, for TACREV dataset in Table 19, and for ACE05 datasets in Table 20.

Table 15: Demo examples used in FewNERD dataset. We color the entity in blue and the analysis in red.

Instruct: Read following sentences and identify what is the entity type of 392 quoted by <t>.
Sentence: Powell v . Texas , <t> 392 <t> U.S. 514 (1968) , was a United States Supreme Court case that ruled that a Texas statute criminalizing public intoxication did not violate the Eighth Amendment protection against cruel and unusual punishment.
(a) 392 is a legal document, a term or a convention in legal sense.
(b) 392 does not belong to any known entities.
(c) 392 refers to a protest, uprising or revolution event
(d) 392 refers to a government or governmental agency
Analysis: In the context you provided, 392 refers to the volume number in the United States Reports where the Supreme Court's decision in Powell v. Texas can be found. However, 392 itself does not refer to a legal document. So 392 do/does not belong to any known entities.
Answer: (b)

Instruct: Read following sentences and identify what is the entity type of The New Yorker quoted by <t>.
Sentence: In 2004 Gourevitch was assigned to cover the 2004 U.S. presidential election for " <t> The New Yorker <t> ".
(a) The New Yorker does not belong to any known entities.
(b) The New Yorker is a broadcast program.
(c) The New Yorker is a kind of written art.
(d) The New Yorker is a media/newspaper organization.
Analysis: The New Yorker is a well-known American magazine that has been published since 1925, and is primarily known for its long-form journalism, commentary, and satire. It has a reputation for publishing high-quality writing on a wide variety of topics, including politics, culture, and the arts. So The New Yorker is a media/newspaper organization.
Answer: (d)

Instruct: Read following sentence and identify what is the entity type of St. quoted by <t>.
Sentence: The May 1980 eruption of Mount <t> St. <t> Helens in the state of Washington seriously affected both 47th Air Division and 92d Bombardment Wing operations at Fairchild AFB , resulting in dispersal of Fairchild 's B-52 and KC-135 aircraft to various bases while around-the-clock shifts removed the volcanic ash from facilities within the base perimeter. "
(a) St. does not belong to any known entities.
(b) St. is a natural disaster event.
(c) St. is a geographic position about mountain.
Analysis: According to the context, St. is an abbreviation of Saint, used in the name of Mount St. Helens, which is an active volcano in the state of Washington. However, St. itself does not refer to anything. So St. do/does not belong to any known entities.
Answer: (a)

Instruct: Read following sentence and identify what is the entity type of Ridzuan quoted by <t>.
Sentence: <t> Ridzuan <t> was promoted to Harimau Muda A for 2014 season .
(a) Ridzuan does not belong to any known entities.
(b) Ridzuan is a person, but not affiliated with following professions: actor, artist, author, director, politician, scholar, soldier.
(c) Ridzuan is an athlete.
Analysis: The mention of 'Harimau Muda A' indicates that it is a sports-related context, and 'promoted' implies a progression or advancement within the sports team. So Ridzuan is an athlete.
Answer: (c)

Table 16: Demo examples used in TACREV dataset. We color the subject and object entities in blue and the analysis in red.

Instruct: Read the sentence and determine the relation between she and lawyer quoted by <t>.
Sentence: The <t> lawyer <t> denied Italian news reports that she wept while addressing the court, but said Knox was upset as <t> she <t> recounted “ the pressure, the aggressiveness of the police who called her a liar . ”
(a) she is the other family member of lawyer
(b) she is a lawyer
(c) she has no known relations to lawyer
Analysis: In the sentence, the word ’she’ refers to someone who was upset while recounting certain events in court. The word ’lawyer’ refers to someone who denied a news report about that same person weeping in court. There is no information in the sentence to indicate that the two individuals are related in any way. So she has no known relations to lawyer.
Answer: (c)

Instruct: Read the sentence and determine the relation between MEF and Myanmar Equestrian Federation quoted by <t>.
Sentence: YANGON , Dec. 27 -LRB- Xinhua -RRB- – Myanmar will hold a horse race in Yangon to commemorate the country ’s 63rd Anniversary Independence Day , the <t> Myanmar Equestrian Federation <t> -LRB- <t> MEF <t> -RRB- confirmed to Xinhua on Monday.
(a) MEF is also known as Myanmar Equestrian Federation
(b) MEF has political affiliation with Myanmar Equestrian Federation
(c) MEF has no known relations to Myanmar Equestrian Federation
Analysis: The symbols -LRB- and -RRB- in the sentence stand for left and right round brackets and are used to enclose the abbreviation ’MEF’ to indicate that it is a replacement for the longer name ’Myanmar Equestrian Federation. So MEF is also known as Myanmar Equestrian Federation.
Answer: (a)

Instruct: Read the sentence and determine the relation between Douglas Flint and chairman quoted by <t>.
Sentence: At the same time , Chief Financial Officer <t> Douglas Flint <t> will become <t> chairman <t> , succeeding Stephen Green who is leaving to take a government job.
(a) Douglas Flint has no known relations to chairman
(b) Douglas Flint is a chairman
(c) Douglas Flint is the employee of chairman
Analysis: The sentence states that Chief Financial Officer Douglas Flint Douglas Flint will succeed Stephen Green as a chairman. So Douglas Flint is a chairman.
Answer: (b)

Instruct: Read the sentence and determine the relation between FAA and U.S. quoted by <t>.
Sentence: On its Web site , the <t> U.S. <t> <t> FAA <t> says the Category 2 rating means the country lacks the laws or regulations that are needed for the certification and oversight of air carriers , according to minimum international standards.
(a) FAA is also known as U.S.
(b) FAA has no known relations to U.S.
(c) FAA has a headquarter in the country U.S.
Analysis: The sentence states that the FAA says the Category 2 rating means the country lacks the laws or regulations needed for the certification and oversight of air carriers, indicating that the FAA is responsible for overseeing aviation regulations in the country. Actually the FAA (Federal Aviation Administration) is a U.S. government agency responsible for regulating and overseeing civil aviation in the United States, and it has its headquarters in Washington, D.C.. So FAA has a headquarter in the country U.S..
Answer: (c)

Table 17: Demo examples used in ACE05 dataset. We color the trigger word in blue and the analysis in red.

Instruct: Read following sentences and identify what event is triggered by the word loan quoted by <t>.
Sentence: Separately , former WorldCom CEO Bernard Ebbers failed on April 29 to make a first repayment of 25 million dollars, plus interest, on a 400-million - dollar <t> loan <t> from MCI, the Journal said, citing SEC documents.
(a) The word loan does not trigger any known event.
(b) The word loan triggers a TRANSFER-MONEY event: giving, receiving, borrowing, or lending money when it is NOT in the context of purchasing something.
(c) The word loan triggers a DECLARE-BANKRUPTCY event: an Entity officially requests legal protection from debt collection due to an extremely negative balance sheet.
Analysis: In the given sentence, the word loan is used to describe the 400-million-dollar loan from MCI to former WorldCom CEO Bernard Ebbers, which he failed to repay on time. This situation clearly involves a transfer of money without the context of purchasing something, which falls under the TRANSFER-MONEY event. So the word loan triggers a TRANSFER-MONEY event: giving, receiving, borrowing, or lending money when it is NOT in the context of purchasing something.
Answer: (b)

Instruct: Read following sentences and identify what event is triggered by the words treated quoted by <t>.
Sentence: When she 's in Germany , Lynch will be <t> treated <t> for bullet wounds and broken bones .
(a) The word treated triggers an INJURE event: a PERSON gets/got injured whether it occurs accidentally, intentionally or even self-inflicted.
(b) The word treated does not trigger any known event.
(c) The word treated triggers a TRANSPORT event: an ARTIFACT (WEAPON or VEHICLE) or a PERSON is moved from one PLACE (GEOPOLITICAL ENTITY, FACILITY, LOCATION) to another.
Analysis: The sentence suggests that Lynch has already been injured and will receive medical treatment in Germany for her injuries. The word 'treated' simply describes the medical care she will receive and does not indicate a new event or action taking place. So the word treated does not trigger any known event.
Answer: (b)

Instruct: Read following sentences and identify what event is triggered by the words buy quoted by <t>.
Sentence: And I won't dwell on the irony of an Oracle employee being driven out of Oracle , starting his own company , and forcing Ellison to spend $ 10.3 billion to get his company – but not him – back (though it does rather delightfully remind me of Coca - Cola basically giving away the bottling franchise and then spending billions to <t> buy <t> it back) .
(a) The word buy triggers a DECLARE-BANKRUPTCY event: an Entity officially requests legal protection from debt collection due to an extremely negative balance sheet.
(b) The word buy triggers a TRANSFER-OWNERSHIP event: The buying, selling, loaning, borrowing, giving, or receiving of artifacts or organizations by an individual or organization.
(c) The word buy does not trigger any known event.
Analysis: In the given sentence, the word buy is used to describe the action of Oracle spending $10.3 billion to get a company back. This clearly involves the transfer of ownership of the company from one entity to another. So the word buy triggers a TRANSFER-OWNERSHIP event: The buying, selling, loaning, borrowing, giving, or receiving of artifacts or organizations by an individual or organization.
Answer: (b)

Instruct: Read following sentences and identify what event is triggered by the words set quoted by <t>.
Sentence: British forces also began establishing the country's first postwar administration Tuesday, granting a local sheik power to <t> set <t> up an administrative committee representing the groups in the region.
(a) The word set triggers a START-POSITION event: a PERSON elected or appointed begins working for (or changes offices within) an ORGANIZATION or GOVERNMENT.
(b) The word set triggers a START-ORG event: a new ORGANIZATION is created.
(c) The word set does not trigger any known event.
Analysis: The phrase 'set up' specifically implies the creation or establishment of a new organization or entity, rather than simply the word 'set'. So the word set does not trigger any known event.
Answer: (c)

Table 18: Templates for FewNERD dataset, where {ent} is the placeholder for event type.

Entity	Template
no-entity	{ent} do/does not belong to any known entities.
person-artist/author	{ent} is an artist or author.
person-actor	{ent} is an actor.
art-writtenart	{ent} is a kind of writtenart.
person-director	{ent} is a director.
person-other	{ent} is a person, but not affiliated with following professions: actor, artist, athlete, author, director, politician, scholar, soldier.
organization-other	{ent} pertains to an organization that does not fall under the categories of company, educational institution, government, media, political party, religion, sports league, sports team, band or musical group.
organization-company	{ent} is a company
organization-sportsteam	{ent} is a sports team
organization-sportsleague	{ent} is a sports league
product-car	{ent} is a kind of car
event-protest	{ent} refers to a protest, uprising or revolution event
organization-government/governmentagency	{ent} refers to a government or governmental agency
other-biologything	{ent} is a special term about biology / life science.
location-GPE	{ent} is a kind of geopolitical entity
location-other	{ent} is a geographic locaton that does not fall under the categories of geopolitical entity, body of water, island, mountain, park, road, railway and transit.
person-athlete	{ent} is an athlete or coach.
art-broadcastprogram	{ent} is a broadcast program.
product-other	{ent} is a kind of product that does not fall under the categories of airplane, train, ship, car, weapon, food, electronic game and software.
building-other	{ent} is a kind of building that does not fall under the categories of airport, hospital, hotel, library, restaurant, sports facility and theater
product-weapon	{ent} is a kind of weapon.
building-airport	{ent} is an airport.
building-sportsfacility	{ent} is a sports facility building.
person-scholar	{ent} is a scholar.
art-music	{ent} is a music.
event-other	{ent} refers to some event except attack, election, natural disaster, protest, revolution and sports
other-language	{ent} is a kind of human language.
other-chemicalthing	{ent} is some special term about chemical science.
art-film	{ent} is a film.
building-hospital	{ent} is a hospital.
other-law	{ent} is a legal document, a term or a convention in legal sense.
product-airplane	{ent} is kind of airplane product.
location-road/railway/highway/transit	{ent} is a geographic position about roadways, railways, highways or public transit systems.
person-soldier	{ent} is a soldier
location-mountain	{ent} is geographic position about mountain.
organization-education	{ent} is an educational institute/organization.
organization-media/newspaper	{ent} is a media/newspaper organization.

product-software	{ent} is a software product.
location-island	{ent} is geographic position about island.
location-bodiesofwater	{ent} is geographic position situated near a body of water.
building-library	{ent} is a library.
other-astronomything	{ent} is a special term about astronomy.
person-politician	{ent} is a politician or lawyer or judge.
building-hotel	{ent} is a hotel building.
product-game	{ent} is a electronic game product.
other-award	{ent} is a kind of award.
event-sportsevent	{ent} refers to some event related to sports.
organization-showorganization	{ent} is a band or musical organization.
other-educationaldegree	{ent} is a kind of educational degree.
building-theater	{ent} is a theater.
other-disease	{ent} is a kind of disease.
event-election	{ent} is an event about election.
organization-politicalparty	{ent} is a political party/organization.
other-currency	{ent} is a kind of currency.
event-attack/battle/war/militaryconflict	{ent} is an event about attack, battle, war or military conflict.
product-ship	{ent} is a ship.
building-restaurant	{ent} is a restaurant.
other-livingthing	{ent} is a living animal/creature/organism.
art-other	{ent} is a work of art, but not belong to the categories of music, film, written art, broadcast or painting.
event-disaster	{ent} is a natural disaster event.
organization-religion	{ent} is a religious organization.
other-medical	{ent} refers to some kind of medicine.entity
location-park	{ent} is a park.
other-god	{ent} is a god in some legend/religious story.
product-food	{ent} is a kind of food.
product-train	{ent} is a kind of train(vehicle).
art-painting	{ent} is an art painting.

Table 19: Templates for TACREV dataset, where {subj} and {obj} are the placeholders for subject and object entities. Copied from (Lu et al., 2022a)

Relation	Template
no_relation	{subj} has no known relations to {obj}
per:stateorprovince_of_death	{subj} died in the state or province {obj}
per:title	{subj} is a {obj}
org:member_of	{subj} is the member of {obj}
per:other_family	{subj} is the other family member of {obj}
org:country_of_headquarters	{subj} has a headquarter in the country {obj}
org:parents	{subj} has the parent company {obj}
per:stateorprovince_of_birth	{subj} was born in the state or province {obj}
per:spouse	{subj} is the spouse of {obj}
per:origin	{subj} has the nationality {obj}
per:date_of_birth	{subj} has birthday on {obj}
per:schools_attended	{subj} studied in {obj}
org:members	{subj} has the member {obj}
org:founded	{subj} was founded in {obj}
per:stateorprovinces_of_residence	{subj} lives in the state or province {obj}
per:date_of_death	{subj} died in the date {obj}
org:shareholders	{subj} has shares hold in {obj}
org:website	{subj} has the website {obj}
org:subsidiaries	{subj} owns {obj}
per:charges	{subj} is convicted of {obj}
org:dissolved	{subj} dissolved in {obj}
org:stateorprovince_of_headquarters	{subj} has a headquarter in the state or province {obj}
per:country_of_birth	{subj} was born in the country {obj}
per:siblings	{subj} is the siblings of {obj}
org:top_members/employees	{subj} has the high level member {obj}
per:cause_of_death	{subj} died because of {obj}
per:alternate_names	{subj} has the alternate name {obj}
org:number_of_employees/members	{subj} has the number of employees {obj}
per:cities_of_residence	{subj} lives in the city {obj}
org:city_of_headquarters	{subj} has a headquarter in the city {obj}
per:children	{subj} is the parent of {obj}
per:employee_of	{subj} is the employee of {obj}
org:political/religious_affiliation	{subj} has political affiliation with {obj}
per:parents	{subj} has the parent {obj}
per:city_of_birth	{subj} was born in the city {obj}
per:age	{subj} has the age {obj}
per:countries_of_residence	{subj} lives in the country {obj}
org:alternate_names	{subj} is also known as {obj}
per:religion	{subj} has the religion {obj}
per:city_of_death	{subj} died in the city {obj}
per:country_of_death	{subj} died in the country {obj}
org:founded_by	{subj} was founded by {obj}

Table 20: Templates for ACE05 dataset, where {evt} is the placeholder for event type.

Event	Template
no-event	The word {evt} does not trigger any known event.
Movement.Transport	The word {evt} triggers a TRANSPORT event: an ARTIFACT (WEAPON or VEHICLE) or a PERSON is moved from one PLACE (GEOPOLITICAL ENTITY, FACILITY, LOCATION) to another.
Personnel.Elect	The word {evt} triggers an ELECT event which implies an election.
Personnel.Start-Position	The word {evt} triggers a START-POSITION event: a PERSON elected or appointed begins working for (or changes offices within) an ORGANIZATION or GOVERNMENT.
Personnel.Nominate	The word {evt} triggers a NOMINATE event: a PERSON is proposed for a position through official channels.
Conflict.Attack	The word {evt} triggers an ATTACK event: a violent physical act causing harm or damage.
Personnel.End-Position	The word {evt} triggers an END-POSITION event: a PERSON stops working for (or changes offices within) an ORGANIZATION or GOVERNMENT.
Contact.Meet	The word {evt} triggers a MEET event: two or more entities come together at a single location and interact with one another face-to-face.
Life.Marry	The word {evt} triggers a MARRY event: two people are married under the legal definition.
Contact.Phone-Write	The word {evt} triggers a PHONE-WRITE event: two or more people directly engage in discussion which does not take place 'face-to-face'.
Transaction.Transfer-Money	The word {evt} triggers a TRANSFER-MONEY event: giving, receiving, borrowing, or lending money when it is NOT in the context of purchasing something.
Justice.Sue	The word {evt} triggers a SUE event: a court proceeding has been initiated for the purposes of determining the liability of a PERSON, ORGANIZATION or GEOPOLITICAL ENTITY accused of committing a crime or neglecting a commitment
Conflict.Demonstrate	The word {evt} triggers a DEMONSTRATE event: a large number of people come together in a public area to protest or demand some sort of official action. For eample: protests, sit-ins, strikes and riots.
Business.End-Org	The word {evt} triggers an END-ORG event: an ORGANIZATION ceases to exist (in other words, goes out of business).
Life.Injure	The word {evt} triggers an INJURE event: a PERSON gets/got injured whether it occurs accidentally, intentionally or even self-inflicted.
Life.Die	The word {evt} triggers a DIE event: a PERSON dies/died whether it occurs accidentally, intentionally or even self-inflicted.
Justice.Arrest-Jail	The word {evt} triggers a ARREST-JAIL event: a PERSON is sent to prison.
Transaction.Transfer-Ownership	The word {evt} triggers a TRANSFER-OWNERSHIP event: The buying, selling, loaning, borrowing, giving, or receiving of artifacts or organizations by an individual or organization.
Justice.Execute	The word {evt} triggers an EXECUTE event: a PERSON is/was executed
Justice.Trial-Hearing	The word {evt} triggers a TRIAL-HEARING event: a court proceeding has been initiated for the purposes of determining the guilt or innocence of a PERSON, ORGANIZATION or GEOPOLITICAL ENTITY accused of committing a crime.
Justice.Sentence	The word {evt} triggers a SENTENCE event: the punishment for the DEFENDANT is issued
Life.Be-Born	The word {evt} triggers a BE-BORN event: a PERSON is given birth to.
Justice.Charge-Indict	The word {evt} triggers a CHARGE-INDICT event: a PERSON, ORGANIZATION or GEOPOLITICAL ENTITY is accused of a crime
Business.Start-Org	The word {evt} triggers a START-ORG event: a new ORGANIZATION is created.
Justice.Convict	The word {evt} trigges a CONVICT event: a PERSON, ORGANIZATION or GEOPOLITICAL ENTITY is convicted whenever it has been found guilty of a CRIME.
Business.Declare-Bankruptcy	The word {evt} triggers a DECLARE-BANKRUPTCY event: an Entity officially requests legal protection from debt collection due to an extremely negative balance sheet.
Justice.Release-Parole	The word {evt} triggers a RELEASE-PAROLE event.

Justice.Fine	The word {evt} triggers a FINE event: a GEOPOLITICAL ENTITY, PERSON or ORGANIZATION get financial punishment typically as a result of court proceedings.
Justice.Pardon	The word {evt} triggers a PARDON event: a head-of-state or their appointed representative lifts a sentence imposed by the judiciary.
Justice.Appeal	The word {evt} triggers a APPEAL event: the decision of a court is taken to a higher court for review
Business.Merge-Org	The word {evt} triggers a MERGE-ORG event: two or more ORGANIZATION Entities come together to form a new ORGANIZATION Entity.
Justice.Extradite	The word {evt} triggers a EXTRADITE event.
Life.Divorce	The word {evt} triggers a DIVORCE event: two people are officially divorced under the legal definition of divorce.
Justice.Acquit	The word {evt} triggers a ACQUIT event: a trial ends but fails to produce a conviction.